Bornes de Risque pour CART en Classification

Servane Gey

Laboratoire MAP5, Université Paris Descartes, 45 rue des Saints Pères, 75270 Paris Cedex 06, France

Résumé

Des bornes de risque pour les arbres de classification CART (Breiman *et. al.* 1984) sont obtenues sous une condition de marge dans le cadre de la classification supervisée binaire. Ces bornes sont prises conditionnellement à la construction de l'arbre de plus grande profondeur construit sur un échantillon d'apprentissage. Elles permettent de valider le choix de la pénalité dans l'algorithme d'élagage d'une part, et de montrer que la sélection d'un sous-arbre dans la suite de sous-arbres élagués à l'aide d'un échantillon-témoin n'altère pas trop la qualité du classificateur sélectionné d'autre part. Dans le cadre de la classification binaire, et sous une condition de marge, ces bornes de risque, obtenues par des techniques de sélection de modèles, permettent de valider l'algorithme CART.

Abstract

Risk bounds for Classification and Regression Trees (CART, Breiman *et. al.* 1984) classifiers are obtained under a margin condition in the binary supervised classification framework. These risk bounds are obtained conditionally on the construction of the maximal binary tree and permit to prove that the linear penalty used in the CART pruning algorithm is valid under a margin condition. It is also shown that, conditionally on the construction of the maximal tree, the final selection by test sample does not alter dramatically the estimation accuracy of the Bayes classifier. In the two-class classification framework with margin condition, the risk bounds that are proved, obtained by using penalized model selection, validate the CART algorithm which is used in many data mining applications such as Biology, Medicine or Image Coding.

Mots-clés : Apprentissage et classification, Statistique mathématique.

1 Introduction

Les arbres de décision CART (Classification and Regression Trees), proposés par Breiman, Friedman, Olshen et Stone (1984), permettent de construire de manière simple et rapide des classificateurs ou des régresseurs constants par morceaux à partir d'un échantillon d'apprentissage. Cet algorithme est basé sur une découpe diadique récursive de l'espace des covariables qui affecte à chaque partie de la partition ainsi construite une étiquette

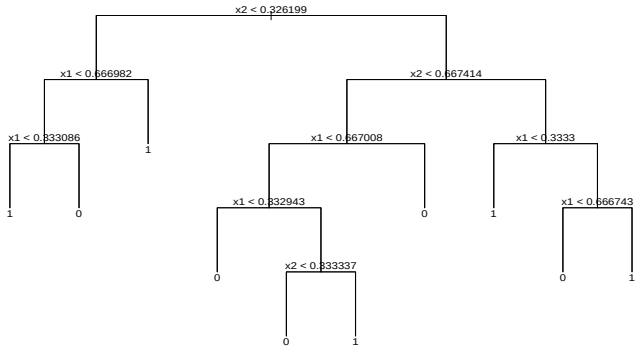

FIG. 1 – Exemple d'arbre de décision

(*classificateurs*) ou une valeur réelle (*régresseurs*). Un exemple simple d'arbre de décision est donné dans la Figure 1. Dans cet exemple, la partition de l'espace des covariables $[0;1]^2$ est définie par la succession de questions posées à chaque nœud de l'arbre : si la réponse est positive, on descend dans le nœud gauche, sinon on descend dans le nœud droit. Ainsi, la première question correspond à une partition en deux sous-ensembles de l'espace des covariables. Puis, chaque question posée à chaque nœud sépare en deux sous-parties filles la partie parente correspondant au nœud en question. Une fois la partition obtenue par découpe diadique récursive, on associe à chaque partie une valeur pour la variable à prédire, donnée par la valeur associée à la feuille de l'arbre correspondante.

Nous nous intéressons plus particulièrement à CART en classification, c'est-à-dire lorsque la variable expliquée est qualitative. Nous verrons dans la suite que CART peut être vu comme une procédure de sélection de modèles, où la collection de modèles considérée est constituée d'arbres de décision construits sur l'échantillon d'apprentissage. Nous rappellerons également comment l'algorithme sélectionne un petit nombre d'arbres dans cette collection par le biais d'un critère pénalisé, dont la fonction de pénalité est proportionnelle au nombre de feuilles des arbres considérés. Un arbre est ensuite choisi dans cette petite collection à l'aide d'un échantillon-témoin. Nous exhiberons des bornes de risques permettant, sous certaines conditions sur la distribution des variables, de valider le choix de la forme de la pénalité dans le critère d'élagage, et également de montrer que la perte occasionnée par la sélection via échantillon-témoin est convenablement contrôlée.

La procédure CART s'inscrit dans le cadre plus général suivant. On considère $\mathcal{L} = \{(X_1, Y_1), \ldots, (X_N, Y_N)\}$ une réalisation de N copies indépendantes du couple de variables (X, Y) de loi commune P, où la variable explicative X prend ses valeurs dans un espace

mesurable $\mathcal{X}$ et est associée à une étiquette Y à valeurs dans $\{0;1\}$. On cherche à expliquer Y en fonction de X, c'est-à-dire on cherche un classificateur f de $\mathcal{X}$ dans $\{0;1\}$ qui rende minimale l'erreur de classification définie par

$$\mathrm{P}(f(X) \neq Y).$$

Si P était connue, trouver un tel classificateur serait aisé. Il suffirait de prendre (voir Devroye, Györfi et Lugosi (1996)) le classificateur de Bayes défini pour tout $x \in \mathcal{X}$ par

$$f^*(x) = \mathbb{1}_{\eta(x)>1/2}, \tag{1}$$

où $\eta(x)$ est l'espérance conditionnelle de Y sachant $X = x$, en d'autres termes

$$\eta(x) = \mathrm{P}[Y = 1 \mid X = x]. \tag{2}$$

P étant inconnue, on cherche à construire sur $\mathcal{L}$ un classificateur $\hat{f}$ qui sera aussi proche que possible de f^* au sens de la perte définie par

$$l(\hat{f}, f^*) = \mathrm{P}(\hat{f}(X) \neq Y) - \mathrm{P}(f^*(X) \neq Y), \tag{3}$$

où $\hat{f}$ est considéré ici comme une fonctionnelle de X. f^* minimisant $\mathrm{P}(f(X) \neq Y)$ sur l'ensemble des classificateurs, $l(., f^*)$ est positive. Le risque d'un classificateur $\hat{f}$ sera alors la perte moyenne $\mathbb{E}[l(\hat{f}, f^*)]$ calculée par rapport à la mesure produit sur $\mathcal{L}$.

Afin de s'assurer que les observations sont suffisamment bien distribuées sous la loi P inconnue, nous ferons l'hypothèse de marge suivante : il existe une constante $h \in [0;1]$ telle que

$$\mathrm{P}\left(|2\eta(X) - 1| \leqslant h\right) = 0, \tag{4}$$

où η est définie par (2). L'hypothèse (4) a permis à Massart et Nédélec (2006) d'améliorer les bornes de risque obtenues dans un premier temps par Vapnik (voir Devroye, Györfi et Lugosi (1996)) pour un classificateur construit sur un modèle. Il s'agit d'un cas particulier de l'hypothèse de marge proposée par Mammen et Tsybakov (1999), qui permet d'assurer que les observations sont suffisamment bien distribuées pour être classables.

Dans toute la suite, pour un arbre binaire T, nous noterons $\widetilde{T}$ l'ensemble des parties définies par les feuilles de T, et $|\widetilde{T}|$ son cardinal. Le modèle $\mathcal{F}_T$ associé à T sera alors défini par

$$\mathcal{F}_T = \left\{ f :\ \mathcal{X} \to\ \{0;1\}\ ;\ f = \sum_{t \in \widetilde{T}} f_t \mathbb{1}_t,\ (f_t)_{t \in \widetilde{T}} \in \{0;1\}^{|\widetilde{T}|} \right\} \tag{5}$$

où $\mathbb{1}_t$ est l'indicatrice de la partie t. L'algorithme CART consistera alors à choisir un modèle dans une collection $(\mathcal{F}_T)_{T \in \mathcal{M}_N}$ de modèles aléatoires, et à estimer le classificateur de Bayes (1) sur ce modèle.

2 CART en Classification

Supposons que l'on dispose de l'échantillon $\mathcal{L}$, que l'on sépare en deux sous-échantillons $\mathcal{L}_1$ (*échantillon d'apprentissage*) de taille n_1, et $\mathcal{L}_2$ (*échantillon-témoin*) de taille n_2, avec $n_1 + n_2 = N$. L'algorithme CART peut alors être décomposé en trois étapes :

E1 Construction d'un arbre binaire T_{max} de grande profondeur par découpe diadique récursive de l'espace $\mathcal{X}$ (arbre maximal) sur $\mathcal{L}_1$.

E2 Extraction d'une suite $(T_k)_{1 \leqslant k \leqslant K}$ d'arbres emboîtés les uns dans les autres et issus de T_{max} (élagage) par le biais de $\mathcal{L}_1$.

E3 Sélection d'un classificateur dans la suite extraite à l'étape **E2** par le biais de $\mathcal{L}_2$.

L'étape **E1** est récursive, et basée sur la minimisation d'une fonction convexe des proportions de chaque étiquette dans les nœuds fils du nœud considéré. Les résultats que nous présenterons étant conditionnels à cette étape, nous ne la détaillerons pas ici.

Etape **E2** *: Elagage*

Pour un arbre T fixé, on dit que T' est élagué de T si T' est un sous-arbre binaire de T ayant même racine. On notera cette relation $T' \preceq T$. Il s'agit d'une relation d'ordre sur l'ensemble des sous-arbres élagués de T_{max}.
La collection de modèles aléatoires $\mathcal{M}_N$ est alors définie comme l'ensemble de tous les sous-arbres élagués de l'arbre maximal T_{max}. Cette collection ayant une complexité exponentielle en n_1, il est en général impossible de visiter tous les sous-arbres pour en choisir un par le biais de $\mathcal{L}_2$. La procédure d'élagage permet donc, dans un premier temps, de réduire de façon drastique la collection de sous-arbres à considérer. Cette procédure est basée sur le critère pénalisé construit de la manière suivante.
Pour un arbre T élagué de T_{max}, on construit le classificateur $\hat{f}_T$ minimisant le critère empirique $\gamma_{n_1}(f) = n_1^{-1} \sum_{\{(X_i,Y_i) \in \mathcal{L}_1\}} \mathbb{1}_{f(X_i) \neq Y_i}$ sur le modèle $\mathcal{F}_T$ défini par (5) :

$$\hat{f}_T \quad = \quad \sum_{t \in \widetilde{T}} \text{argmax}_{\{Y_i \ ; \ X_i \in t\}} |\{Y_i \ ; \ X_i \in t\}| \mathbb{1}_t.$$

Le critère pénalisé est alors défini pour toute température $\alpha > 0$ et tout $T \preceq T_{max}$ par

$$\text{crit}_\alpha(T) \quad = \quad \gamma_{n_1}(\hat{f}_T) + \alpha \frac{|\widetilde{T}|}{n_1}. \tag{6}$$

Pour tout $\alpha > 0$, il existe un unique sous-arbre T_α minimisant crit_α et élagué de tout les sous-arbres minimisant ce même critère (voir Breiman, Friedman, Oshen et Stone (1984)). L'élagage consiste à couper de manière récursive les branches de T_{max} en faisant augmenter la température dans le critère pénalisé. Cette procédure peut être résumée par le théorème suivant :

Théorème (Breiman, Friedman, Olshen, Stone)
Il existe une suite strictement croissante de températures $0 = \alpha_1 < \ldots < \alpha_K$ *associée à une suite strictement décroissante de sous-arbres élagués les uns des autres* $T_1 \succ \ldots \succ T_K$ *telles que, pour tout* $k \in \{1; \ldots; K\}$ *et tout* $\alpha \in [\alpha_k; \alpha_{k+1}[$, $T_\alpha = T_{\alpha_k} = T_k$.

La suite $(T_k)_{k\in\{1;\ldots;K\}}$ est donc convenablement choisie suivant le critère fixé crit_α dans le sens où elle contient toute l'information statistique. Les bornes de risque que nous obtenons permettent de montrer que ce critère est convenablement choisi sous l'hypothèse de marge. Il s'agit ensuite de sélectionner un arbre dans la suite $(T_k)_{1\leqslant k\leqslant K}$ par le biais de l'échantillon-témoin $\mathcal{L}_2$.

Etape **E3** *: Sélection finale*

Etant-donnée la suite de sous-arbres élagués $(T_k)_{k\in\{1;\ldots;K\}}$, on dispose d'une collection de classificateurs $(\hat{f}_{T_k})_{k\in\{1;\ldots;K\}}$. Un modèle d'arbre de décision est alors choisi par le biais de l'échantillon $\mathcal{L}_2$ en minimisant le critère empirique $\gamma_{n_2}(f) = n_2^{-1} \sum_{\{(X_i,Y_i)\in\mathcal{L}_2\}} \mathbb{1}_{f(X_i)\neq Y_i}$:

$$\hat{k} \;=\; \text{argmin}_{k\in\{1;\ldots,K\}} \gamma_{n_2}(\hat{f}_{T_k}).$$

Le classificateur final sera alors

$$\tilde{f} \;=\; \hat{f}_{T_{\hat{k}}} \tag{7}$$

3 Bornes de Risque

Etant donné que la collection de modèles $\mathcal{M}_N$ dépend de l'arbre maximal aléatoire T_{max}, les bornes de risque que nous proposons pour le classificateur $\tilde{f}$ défini par (7) sont prises conditionnellement à T_{max}. Partant de la perte (3), le risque conditionnel de $\tilde{f}$ est alors défini par :

$$R_{\mathcal{L}_1}(\tilde{f}) = \mathbb{E}[l_{n_1}(\tilde{f}, f^*) \mid \mathcal{L}_1],$$

où l_{n_1} est la perte définie conditionnellement à la grille $\{X_i \,;\, (X_i, Y_i) \in \mathcal{L}_1\}$, soit pour un classificateur g

$$l_{n_1}(g, f^*) = \mathbb{E}_Y \left[\frac{1}{n_1} \sum_{\{X_i \,;\, (X_i,Y_i)\in\mathcal{L}_1\}} \mathbb{1}_{g(X_i)\neq Y_i} - \mathbb{1}_{f^*(X_i)\neq Y_i} \right],$$

où $\mathbb{E}_Y$ est l'espérance prise sous la loi marginale de Y. Nous obtenons le résultat suivant :

Théorème 1. *Soit* V *la dimension de Vapnik-Chervonenkis de l'ensemble des parties utilisées pour partitionner* $\mathcal{X}$ *lors de l'étape* **E1**. *On suppose que* $n_1 > V$.

Sous l'hypothèse de marge (4), il existe des constantes universelles $C_1 > 1$, $C_2 > 0$ et $C_3 > 0$ telles que

$$R_{\mathcal{L}_1}(\tilde{f}) \leqslant C_1 \inf_{T \preceq T_{max}} \left[\inf_{g \in \mathcal{F}_T} l_{n_1}(g, f^*) + h^{-1} \log\left(\frac{n_1}{V}\right) \frac{|\widetilde{T}|}{n_1} \right] + h^{-1} \frac{C_2}{n_1} + C_3 \; h^{-1} \frac{\log K}{n_2}$$

La première partie de la borne permet de valider la forme de la pénalité utilisée dans le critère d'élagage, linéaire en le nombre de feuilles. La deuxième partie permet d'assurer que l'utilisation d'un échantillon-témoin n'altère pas trop la qualité du classificateur final $\tilde{f}$ construit par CART. Le fait que la marge inconnue apparaisse dans la fonction de pénalité confirme le fait que la température à choisir dans le critère doit être estimée empiriquement, ce qui est fait ici en utilisant $\mathcal{L}_2$.

Cette borne de risque peut être détaillée par les deux bornes suivantes, correspondant à l'élagage d'une part, et à la sélection par échantillon-témoin d'autre part :

Proposition 1. *Pour $\alpha > 0$, soit T_α le plus petit sous-arbre de T_{max} minimisant* crit_α *défini par (6). Soit $P_{\mathcal{L}_2}$ la mesure produit sur $\mathcal{L}_2$ et $\alpha_{n_1,V} = 2 + V/2(1 + \log(n_1/V))$. Sous l'hypothèse de marge (4), il existe une température $\alpha_0 > 0$ telle que, pour tout $\alpha > \alpha_0 h^{-1} \alpha_{n_1,V}$, il existe Σ_α, $C_\alpha > \alpha_0$ et C' telles que*

$$l_{n_1}(\hat{f}_{T_\alpha}, f^*) \leqslant C_\alpha \inf_{T \preceq T_{max}} \left[\inf_{g \in \mathcal{F}_T} l_{n_1}(g, f^*) + \alpha \frac{|\widetilde{T}|}{n_1} \right] + C' h^{-1} \frac{1 + \xi}{n_1}$$

sur un espace Ω_ξ tel que $\mathrm{P}_{\mathcal{L}_2}(\Omega_\xi) \geq 1 - 2\Sigma_\alpha e^{-\xi}$.

Remarque : Σ_α et C_α sont deux fonctions croissantes de α.

Proposition 2. *Soit $\tilde{f}$ défini par (7). Sous l'hypothèse de marge (4), il existe trois constantes universelles $C > 1$, $C' > 3/2$ et $C'' > 3/2$ telles que*

$$R_{\mathcal{L}_1}(\tilde{f}) \leqslant C \inf_{1 \leqslant k \leqslant K} l_{n_1}(\hat{f}_{T_k}, f^*) + C' \; h^{-1} \frac{\log K}{n_2} + h^{-1} \frac{C''}{n_2}$$

Bibliographie

[1] Breiman L., Friedman J., Olshen R. et Stone, C. (1984) *Classification And Regression Trees*, Chapman & Hall.

[2] Devroye L., Györfi L. et Lugosi G. (1996) *A Probabilistic Theory of Pattern Recognition*, Applications of Mathematics (New York), Springer-Verlag, 31.

[3] Mammen E. et Tsybakov A. (1999) *Smooth Discrimination Analysis*, Annals of Statistics, 6, 1808–1829.

[4] Massart P. et Nédélec E. (2006) *Risk Bounds for Statistical Learning*, Annals of Statistics, 34, 2326–2366.